A Kernel Model with Conditional Moving Windows for the Prediction of Transmembrane Helices in Proteins

Massimo Mucciardi[1], Cinzia Di Salvo[2], Giovanni Pirrotta[3]

[1]Department of Economics, Statistics, Mathematics e Sociology, University of Messina, Italy

[2]Department Animal Biology and Marine Ecology, University of Messina, Italy

[3]Department of Social and Human Sciences, University of Messina, Italy

Emails: [1]massimo.mucciardi@unime.it; [2]cdisalvo@unime.it; [3]gpirrotta@unime.it

Abstract

The determination of some proteins structure at high resolution often results difficult from an experimental point of view. It is true especially for many integral membrane proteins and intrinsically disordered proteins. In this context, we try to develop a statistical technique alternative to the classical methods used to calculate hydropathy graphics, called Kernel Windows Method with Conditional Moving Windows (KWMCMW); the latter has been implemented in order to improve the degree of accuracy of the transmembrane helices (TMH) prediction. The KWMCMW introduces the effect of distance in calculating the hydropathy value for each amino acid (AA) residue in the sequence. Consequently, this new method is extended to a peripheral membrane protein with partially unknown structure. The implementation of specific software written in Python language makes the method suitable for application to any membrane protein.

Keywords

Kernel Function; Transmembrane Helices Prediction; Hydropathy Scale; Conditional Moving Windows; Python Language

Introduction

Computational algorithms in conjunction with statistical analysis are useful methods to identify transmembrane helices (TMH) in a protein, since many of these remain recalcitrant to structure determination by X-ray crystallography. So different methods have been developed in order to recognize regions of the proteins that can be ascribed to α-helix, β-strands or random coil [1-2]. Generally, secondary structure prediction algorithms can be categorized into three main classes: physicochemical methods based on various hydropathy scales, statistical methods, and machine learning methods based on various learning algorithms such as Neural Networks, Hidden Markov Models, and Support Vector Machines. As regard the physicochemical methods, one of the first approaches is provided by Kyte and Doolittle (KD) [3]. This method, through the study of the graphics of hydropathy, gives a suitable accuracy in the determination of TM helices, and is based on a scale where the hydrophilic and hydrophobic properties of each of the 20 AA side-chains are taken into consideration. The method uses a moving (or sliding) window (MW) that continuously determines the average of the hydrophaty within a segment of predetermined length L as it advances through the sequence. Recently different prediction programs for TM segments have been developed. Among these, Yizhou Li et. al. [4] in order to identify the signal peptide of a protein or predict its cleavage site, implement an automated method by using artificial neural network. In doing this, hydrophaty scale of KD is adopted to code each AA. In fact is reported that the cleavage site has relationship with the neighboring sequence environment, i.e., hydrophobic core h-region.

Based on the selected physicochemical properties, in this paper we show an advance of the Kernel Windows Method (KWM) [5] in order to improve the degree of accuracy of the prediction of TM helices. The method combines the chemical properties and statistical methodology so as to keep into account the proximity between the AA residues along protein sequence. Consequently, we made an application of the method on Myelin Basic Protein (MBP) which belongs to the family of intrinsically disordered proteins, which fail to form rigid 3-D structures under physiological conditions. In fact, although several structural and computational studies [6-7-8] have

elucidated on the secondary structure of the protein, no evidences on the tertiary structure have yet been obtained. Moreover, intrinsically disordered proteins with nonpolar residues can also penetrate the interfacial region of the membrane and reach the hydrophobic core, especially when such proteins, as MBP, are cationic and interact with negatively charged membrane [9]. For these reasons, we use the KWM in order to show if the method could be applied as identifier of the secondary structure element, in particular α-helices interacting with membrane, in MBP. The paper is organized as follows. Section 2 presents the KWM and the advances introduced, section 3 shows the main results obtained using the KWM considering the database of 92 proteins used and the results of the prediction on MBP, finally in the appendix we provide the code used to implement the KWM algorithm.

The Kernel Windows Method (Kwm): Methodology and Advances

Before introducing the solution proposed for the prediction of secondary structure in TM, we should briefly summarize the KWM method [5]. The term 'kernel' refers to non-parametric methods that involve calculations using a well-defined local neighbourhood. So, the kernel is utilized as smoothing filter that keeps in count the distance between AA residues along protein sequence: we assume that the values of hydropathy in the neighbourhood of the central residue are more influential than distant ones. Therefore, the KWM is different to the MW method where all residues have the same weight (1/L) with respect to the central residue. From this point of view, KWM can be seen as a special case of MW. Other features are: a) KWM uses a single sequence as input; b) KWM is a non-parametric model since it does not assume a specific pattern for AA residues in the membrane. So under this assumption, the weights are chosen in order to assign greater importance to nearest residues by using a univariate symmetric kernel function. Let 'L' be the length of the MW and 'i' the position of the residue; then we define the value of kernel hydropathy (Δi) for the i-th residue as:

$$\Delta_i = \sum_{j=-\left(\frac{L-1}{2}\right)}^{\left(\frac{L-1}{2}\right)} H_j K_j \qquad (1)$$

For the j-th residue within the window L, H_j is the value of hydrophaty and K_j is the value of the kernel function. Under this aspect we can observe that KWM works as a two-stage system. First the length (L) of the window is considered and then the kernel function is applied.

The value K_j is determined according to a distance decay model as follows:

$$K_j = \frac{f(|j-i|)}{A} \qquad (2)$$

In this case the distance $|j-i|$ in **Eq.2** is calculated as the difference (in absolute value) between the position of the residue 'j' and the position of the central residue of reference 'i' within the width of the MW chosen. Currently the functions $f(|j-i|)$ considered and implemented are:

1) Gaussian distance decay function;

$$f(|j-i|) = e^{-\varphi\left(\frac{|j-i|}{h}\right)^{\alpha}} \qquad (3)$$

2) Bi-square distance decay function;

$$f(|j-i|) = \left(1-\left(\frac{|j-i|}{h}\right)^2\right)^2 \qquad (4)$$

and Epanechnikov function;

$$f(|j-i|) = \frac{3}{4}\left(1-\left(\frac{|j-i|}{h}\right)^2\right) \qquad (5)$$

In Gaussian function ϕ is the decay parameter, α is a smoothing parameter; usually these parameters are set with ϕ =0.5 and α =2. For all three models h is a non negative parameter, called bandwidth, that controls the width of the kernel [10]. Therefore it is reasonable to assume h<L. A small bandwidth will produce a sharp kernel with many variations in the weights, while a high bandwidth will produce a flat kernel with little variation in the weights. In this last case it is easy to verify that the KWM tends to MW. Note that for the above models when i=j (central residue), the numerator of K_j assumes the value equal to 1 for Gauss and Bi-square functions and 0.75 for Epanecnikov function. Moreover, for all functions, in

order to ensure that the sum of all weights is equal to 1, we divide by the constant A, where A is equal to:

$$A = \sum_{j=-\left(\frac{L-1}{2}\right)}^{\left(\frac{L-1}{2}\right)} f\left(\left|j-i\right|\right) \quad (6)$$

and so for the i-th residue

$$\sum_{j=-\left(\frac{L-1}{2}\right)}^{\left(\frac{L-1}{2}\right)} K_j = 1 \quad (7)$$

.

From this point of view, this kernel is also called 'order 0 kernel' [11]. In addition, for the first and the last residues of the sequence the kernel is automatically calculated with smaller windows. Despite the KWM method can work with any value of the window size (L), in this work we use a model with a conditional windows (KWMCMW) that vary between 11 AA and 27 AA of length to make it compatible with the number of residues required to span the membrane in a helical conformation [1]. Under this condition, the algorithm implemented allows to find the best values of L, h and threshold (Th_I) that optimize a special parameter of 'smooth'[1]. Next, other two thresholds are calculated in order to maximize the values of F-measure and MCC [12]. We call these threshold values Th_F and Th_M respectively (see appendix for more details).

Principal Results

In order to compare the performance of KWMCMW method with a classical MW method (reference model), the same training dataset proteins has been used. We select 92 non redundant protein sequences with experimentally determined transmembrane topology from 3D_helix MPtopo database [13]. Although we used various types of kernel, in this research we present only the results obtained with the Gaussian KWMCMW setting ϕ=0.5 and α=2 (table 1). For the MW model (table 2) we use the same software of the KWMCMW model but with a fixed window equal to 19 AA and a fixed hydropathy threshold greater than 1.6 [1]. For the hydropathy scale of each residue we use the one provided by Kyte and Doolittle [3]. The performance of classification is assessed by these indices: accuracy (ACC), precision (PR), sensitivity (SN, also called the recall), specificity (SP), the correlation coefficient of Matthew (MCC) and F-measure (F-m) [12]. These indices are calculated using the number of true positives (TP), false positives (FP), true negatives (TN) and false negatives (FN) for each classifier. In addition other two indices have been calculated: Coverage (CVR), the ratio between the number of predicted residues and known residues inside the helices; Helix rate (H_r), the ratio between the number of predicted helices and known helices. As can be seen from the comparison of the two tables (table 1 and 2), the KWMCMW method gives better results in all performance indices. Note that this new method provides a low number of false positive compared to the classic method MW.

As mentioned above, in KWMCMW the influence between residues is measured by weighting system using the functions above mentioned. So, to verify that KWMCMW may be applied for the prediction of secondary elements such as α-helices in protein MBP (Human Isoform 5 identifier code P02686-5), we select from the results obtained for the model KWMCMW (table 1) only those proteins (N=34) compatible with the length of the protein MBP (Lp = 170 AA). In this way we obtain a value of L = 15.91, h = 3.59, Th_I=1.10, Th_F = 0.03 and Th_M=0.59. The application of these parameters to KWMCMW (with Th_F=0.03) provides the results shown in fig 1. Consequently, when MW method is applied to the protein, no peaks above the threshold appear (fig. 1). On the contrary, as can be seen in fig. 1, the shape of the KWMCMW evidences the presence of different peaks. As mentioned above, no X ray resolution is yet available for MBP, but divergent results have been obtained from prediction studies or structural ones. In our analysis the KWMCMW method used agrees with recent literature in the prediction of three peaks in particular: the first peak revealed at residues 38-43 is in accordance with the presence of an anphyphatix α-helix, in which the two sets of phe-phe pairs at the position 42-43 are at least partially buried and immobilized in the membrane as revealed by solid–state NMR spectroscopy studies [14-15].

[1] In the first experimental version of the KWM [5] the values of L and h, were obtained by minimizing the coefficient of variation (CV) of the hydrophobicity inside the α-helices while the threshold was fixed at 0.

TABLE 1 STATISTICS FOR KWMCMW (92 PROTEINS)

Statistics	L	h	Th_I	Max_F	Th_F	Max_	Th_	ACC	PR	SN	SP	CVR	H_r
N	92	92	92	92	92	92	92	92	92	92	92	92	92
Mean	17.74	3.97	1.37	0.84	0.22	0.69	0.71	77.14	99.13	52.67	99.83	53.31	0.99
Median	17.00	3.80	1.36	0.85	0.38	0.69	0.76	79.34	100.00	50.61	100.00	52.47	1.00
Mode	11.00	2.90	-0.34	1.00	-1.34	1.00	0.80	100.00	100.00	100.00	100.00	100.00	1.00
Std. Dev.	5.384	1.67	0.62	0.10	0.81	0.17	0.63	17.13	5.38	23.48	0.78	23.74	0.37
Min	11.00	1.30	-0.34	0.46	-2.29	0.24	-0.92	16.46	48.84	4.35	93.99	4.35	0.67
Max	27.00	12.20	2.81	1.00	1.88	1.00	2.77	100.00	100.00	100.00	100.00	100.00	4.50
25° Perc.	13.00	2.90	1.03	0.79	-0.26	0.44	0.36	64.81	100.00	38.06	100.00	38.34	1.00
75° Perc	23.00	4.90	1.73	0.90	0.77	0.75	1.13	91.07	100.00	67.03	100.00	68.22	1.00

Total know helices =310 Total predicted helices =326 Max_F = Max F-Measure Max_M = Max MCC

TABLE 2 STATISTICS FOR MW (92 PROTEINS)

Statistics	L	h	Fm	Th	MCC	Th_M	ACC	PR	SN	SP	CVR	H_r
N	92	92	92	92	92	92	92	92	92	92	92	92
Mean	19.00	-	0.42	1.60	0.34	0.50	65.60	88.71	30.33	95.91	33.11	1.54
Median	19.00	-	0.43	1.60	0.32	0.57	65.33	95.96	27.43	99.63	28.82	1.41
Mode	19.00	-	0.80	1.60	0.00	0.55	58.14	100.00	0.00	100.00	66.67	2.00
Std. Dev.	.00	-	0.22	.00	0.19	0.59	17.24	19.40	20.48	9.62	22.23	0.85
Min	19.00	-	0.00	1.60	0.00	-2.17	26.14	-1.00	0.00	42.86	0.00	0.00
Max	19.00	-	0.93	1.60	0.80	1.53	94.57	100.00	100.00	100.00	116.00	5.00
25° Perc.	19.00	-	0.27	1.60	0.21	0.27	51.18	87.95	15.97	96.01	18.24	1.00
75° Perc	19.00	-	0.59	1.60	0.49	0.86	81.31	100.00	44.16	100.00	46.34	2.00

Total know helices =310 Total predicted helices =421

The peak centred at residues 85-93 corresponds to an α-helix immersed in the bilayer identified by studies of electronic paramagnetic resonance (EPR) [16]. The third peak (148-155) corresponds to the α-helix Calmodulin binding domain [15]. However, apart from these results, we obtain a better definition of the shape respect to the MW method.

Conclusions

From these results we can assert that the KWMCMW, giving a different weight on the hydropathy of residues according to the distance-decay function, well identifies helices in membrane proteins, so it can be considered an alternative method to other existing bioinformatics and statistical methods. Moreover, it also gives useful information on some secondary elements identified as α-helices of MBP, which interact with oligodendrocyte membrane. So, our prediction is in agreement with the presence of hydrophobic α-helices in MBP, as revealed by the recent structural analyses. Further studies could try to apply KWMCMW on others intrinsically disordered proteins with the aim to identify these secondary elements as well. From a statistical point of view, it is evident that it may be necessary to improve the current criterion for estimating the threshold and to implement new kernel functions combining them with other scales of hydropathy. However, further progress will require more learning set sequences. Some researches of this kind are currently in progress by authors.

Appendix: Implementation of the Kwmcmw Algorithm

The KWMCMW algorithm target is to find the best protein threshold able to locate all the segment helices in the protein implementing the kernel function of the **Eq.1**. The algorithm, implemented in Python 2.7 [17] using Scipy and Numpy external scientific libraries, works as follows.

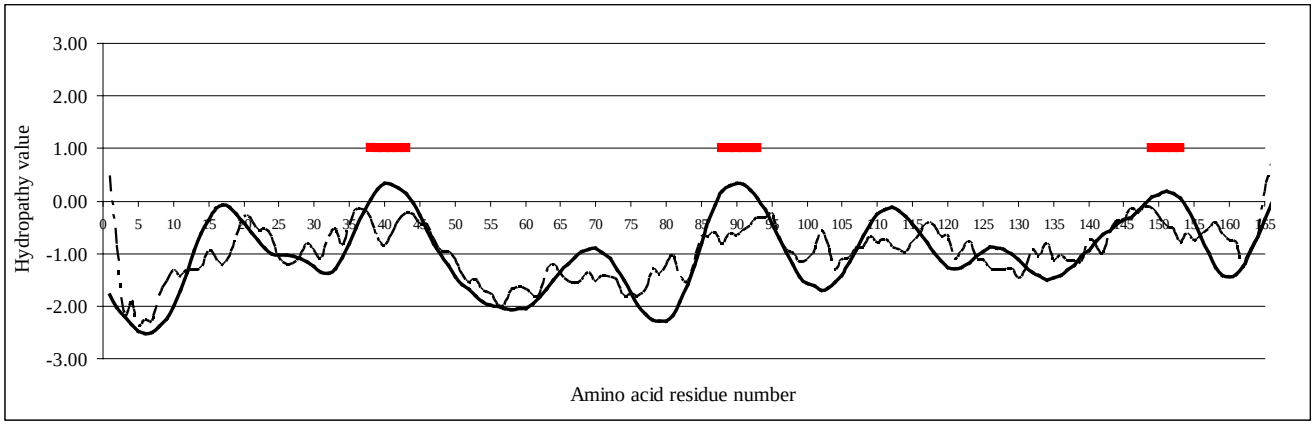

FIG. 1 TM HELICES PREDICTION PROFILES OF HYDROPHOBICITY FOR MBP: SOLID LINE GAUSSIAN KWMCMW; DASHED LINE CLASSICAL MW.THE THREE RED LINES AT THE TOP INDICATE THE POSITIONS OF PREDICTED A-HELICES WITH THE GAUSSIAN KWMCMW

Primary, the algorithm loads the sequences of proteins and the relative transmembrane topology. Second, the algorithm defines the kernel type and the hydrophobic scale: Hydrophobicity, Helicity, KD, GES, and Eisenberg [18]. Only for Gaussian kernel, the alpha and phi values are calculated. Third the algorithm generates all possible combinations of L size (between 11 AA and 27 AA) and h (0<h<L) producing kernel values of hydropathy. Knowing a-priori the transmembrane topology, the algorithm calculates, for each helix (H) and no-helix (NH) segment, some statistics (Mean, Median, Mode, SD, Max, Min and Quartile). So, we consider the following initial threshold: Th*=Max[Max(NH)]. At this point, in order to ensure a very high percentage of precision in the segment classification, the following condition (filter) is adopted: Th*<Min[Max(H)]. This condition creates a restriction in the number of initial combinations of L and h and is essential because it eliminates all those values of hydrophobicity present in the no-helix segments that exceed the values of hydrophobicity in the helix segments. Only for the values of L and h that produce hydrophobic kernel values that exceed the condition, the algorithm computes a special parameter called 'irregularity' to estimate the degree of 'smooth' of kernel. As can be observed in the routine below, this parameter is initially set to a value intentionally large (Irr = 1000). Cycling for all L and h values, the algorithm finds the better threshold (Th_I) that minimizes the 'irregularity' property. In addition, starting from the lower kernel value up to threshold found (Th_I), the algorithm calculates further two new threshold values considering and maximizing MCC and F measures. Therefore, the final output of the algorithm returns 3 threshold values (Th_I, Th_F and Th_M) which can be used to make predictions (see fig. 2 and 3). Currently the library is in an alpha version while the beta version will be released as soon as possible.

```
FOUND = FALSE
MIN_IRR = 1000
LOAD Protein
FOR i=11 TO 27 STEP 2 DO:
  FOR j=1 TO i STEP 0.1 DO:
generator=GaussianKernelGenerator(proteine, scale, i, j, alpha, phi)
    stat = STATISTICS(generator)
    IF stat [MAX-MAX-NH]<stat [MIN-MAX-H] THEN
      Th* = stat[MAX-MAX-NH]
      FOUND = TRUE
      IRR = GetIrregularity(Th*)
      IF IRR < MIN_IRR THEN
        MIN_IRR = IRR
        L = i
        h = j
        Th_I = Th*
        IF MIN_IRR == 0 THEN
          BREAK
IF FOUND THEN
  NORMALIZE(Th_I)
  Th_M = MAX-MCC-MEASURE(Th_I)
  Th_F = MAX-F-MEASURE(Th_I)
```

FIG. 2 CODE OF THE KWMCMW ALGORITHM

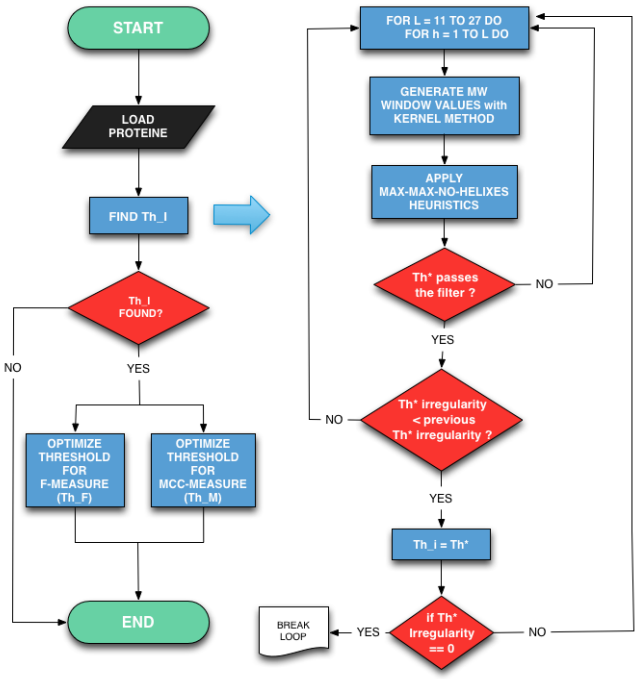

FIG. 3 FLOW CHART OF THE KWMCMW ALGORITHM

REFERENCES

[1] X. Hu, 'Structure Prediction of Membrane Proteins', in Computational Methods for Protein Structure Prediction and Modeling, Editors Y. Xu, D. Xu, H.Liang, pp 65-108, Vol 2, Springer, (2007).

[2] C. P. Chen, A. Kernytsky, and B. Rost, 'Transmembrane helix predictions revisited', Protein Science, 11, (2002).

[3] J. Kyte, R.F. Doolittle, 'A Simple Method for Displaying the Hydropathic Character of a Protein', J. Mol. Biol., vol .157 (1982).

[4] L. Yizhou , W. Zhining, Z. Cuisong, T. Fuyuan, L. Menglong, 'Effects of neighboring sequence environment in predicting cleavage sites of signal peptides', Peptides, vol 29, pp. 1498 – 1504, (2008).

[5] M. Mucciardi, C. Di Salvo, 'A kernel windows method for prediction of protein hydropathy. An application on a myelin associated protein', International Conference on Management Sciences and Information Technology (MSIT 2012), June 1-2, Changsha, China.

[6] R. A. Ridsdale, D. R. Beniac, T. A. Tompkins, M. A. Moscarello, G. Harauz, 'Three-dimensional Structure of Myelin Basic Protein', Journal of Biological Chemistry Vol. 272, No. 7, pp. 4269–4275, (1997).

[7] G. L. Stoner, 'Predicted folding of beta-structure in myelin basic protein', J Neurochem, Aug; 43(2), pp.433-47, (1984).

[8] R. Martenson, 'Prediction of the secondary structure of myelin basic protein, J. Neurochem', Apr 36 (4), pp1543-60, 1981.

[9] J.F. Ellena, J. Moulthrop, J. Wu, M. Rauch, S. Jaysinghne, J.D. Castle, D.S. Cafiso, 'Membrane position of a basic aromatic peptide that sequesters phosphatidylinositol 4,5 bisphosphate determined by site-directed spin labeling and high-resolution NMR'. Biophys J. 87 (5): 3221–3233, (2007).

[10] B.W. Silvermann, 'Density Estimation for Statistics and Data' Analysis, Chapman and Hall, New York, (1986).

[11] A.J. Izenman, 'Recent Developments in Nonparametric Density Estimation', Journal of the American Statistical Association , Vol. 86, No. 413, pp. 205-224, (1991).

[12] Y. Xiong, J. Liu, W. Zhang, T. Zeng, 'Prediction of heme binding residues from protein sequences with integrative sequence profiles', IEEE International Conference on Bioinformatics and Biomedicine, Proteome Science, (2012).

[13] http://blanco.biomol.uci.edu/, S. White Laboratory, (2011).

[14] L. Zhong, V. V. Bamm, M. A.M. Ahmed, 'Solid-state NMR spectroscopy of 18.5 kDa myelin basic protein reconstituted with lipid vesicles: Spectroscopic characterisation and spectral assignments of solvent-exposed protein fragments', Biochimica et Biophysica Acta, pp. 3193–3205, (2007).

[15] R. Dominguez, 'Actin-binding proteins – a unifying hypothesis', Trends Biochem. Sci, Vol.29, 572-5780, (2004).

[16] J. M. Boggs, I. R. Bates, A. A. Musse G. Harauz, 'Interactions Of The 18.5 Kda Myelin Basic protein With Lipid Bilayers: Studies By Electron Paramagnetic Resonance Spectroscopy And Implications For Generation Of Autoimmunity', In Multiple Sclerosis, Myelin Basic Proten, Editor Joan M. Boggs, (2008).

[17] http://www.python.org/, Python Software Foundation, (2012).

[18] C. M. Deber, C. Wang, L. L.P. Liu, A S. Prior, S. Agrawal, B. L. Muskat, A. J. Cuticchia, 'TM Finder: A prediction

program for transmembrane protein segments using a combination of hydrophobicity and nonpolar phase helicity scales', Protein Science, vol 10, pp.212-219, (2001).

Author Introduction

Massimo Mucciardi holds a PhD in Statistics from Messina University, Italy. He is an Assistant Professor in the School of Statistics at the University of Messina, Italy. His main research fields include spatial statistics, kernel methods, GIS and survey sample.

Cinzia Di Salvo holds a PhD in Biochemistry from Messina University, Italy. She has a grant at the University of Messina, Italy. Her main research fields regard protein purification and characterization.

Giovanni Pirrotta holds a PhD in Mathematics from Messina University, Italy. He is a software architect at the University of Messina, Department of Social and Human Sciences. His main research field includes Knowledge Representation, Semantic Web and Education/E-Learning Applications.